# PROYECTO FINAL HENRY

TORRES – MONROY – CHINCUINI – OROPEZA - ZEGARRA

## INTRODUCCIÓN

En este documento vas a encontrar un registro sobre el proceso del proyecto final realizado por nuestro grupo. Para ver el proyecto realizado pueden acceder a nuestro repositorio de github: https://github.com/DJChincuini/PF_Reviews-Recommendation_Henry.

McDonald's, una de las principales cadenas de comida rápida a nivel mundial, nos ha contactado ya que se encuentra desarrollando un nuevo sistema de incentivos para sus empleados en el estado de Florida como parte de su programa para la constante mejora del servicio al cliente. Este programa no solo busca reconocer el excelente rendimiento del equipo, sino también evaluar su eficacia en un entorno operativo real. McDonald's tiene la expectativa de que estos incentivos refuercen su posición como líder en la industria y mejoren la experiencia del cliente en sus establecimientos en Florida y en otros lugares.

Nuestra labor va a ser la de evaluar los resultados del programa generando KPI's acordes y dar el feedback correspondiente a nuestro cliente. Para esto se nos han facilitado distintos datasets que contienen información sobre diferentes empresas en Estados Unidos, por lo que diseñamos un informe EDA en donde se encuentre detallada una descripción de los datasets para pasar a generar una infraestructura en la nube para la automatización del proceso de ETL y que, una vez los datos se encuentren limpios procederemos a analizarlos en base a un dashboard que diseñaremos en base a nuestras necesidades.

## NUESTRO EQUIPO

Somos Data Feedback Solutions, un equipo especializado en transformar la voz de tus clientes en insights que ayudarán a llevar a tu empresa al siguiente nivel. Data Feedback Solutions se conforma por los siguientes profesionales.

| Integrante | Rol |
|---|---|
| Milagros Torres | Data Analyst |
| Alejandra Monroy | Data Analyst |
| Dante Chincuini | Data Engineer |
| Rafael Oropeza | ML Engineer |
| Jonathan Zegarra | ML Engineer |

## ALCANCE DEL PROYECTO

Nuestro equipo se va a encargar de convertir las reseñas de los consumidores de McDonald's en insights en pos de una mejora de rendimiento de la atención al cliente y la "experiencia del consumidor" en los locales de esta cadena de comida rápida. Estas reseñas datan de los años 2010 al 2022, en el marco geográfico del estado de Florida, Estados Unidos. En resumen, nuestra labor se puede reducir en tres puntos:

- Conocer el top 3 de locales con mejores reseñas año a año.
- Evaluar de manera efectiva el desempeño de cada local.
- Identificar áreas de mejora.

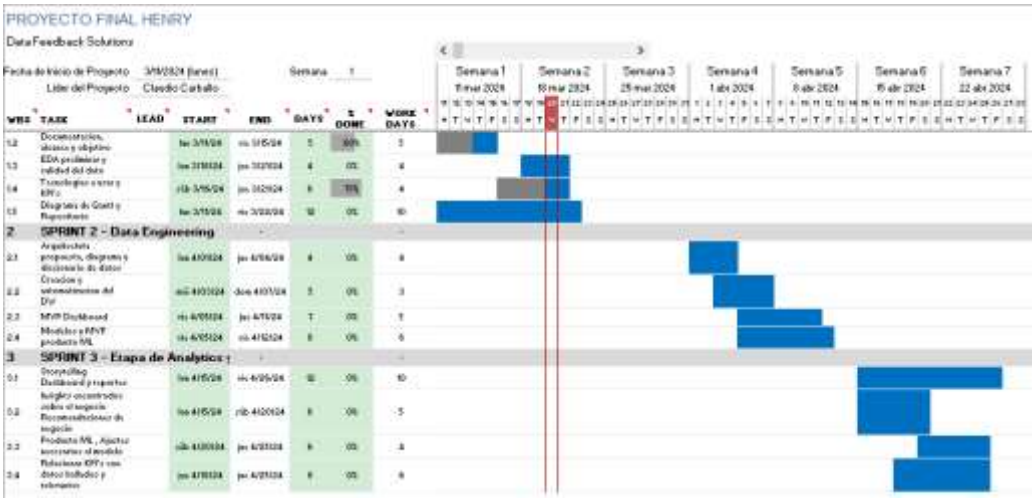
### KPI's

- Elevar en un 5% las reseñas positivas respecto al año anterior.
- Mantener una puntuación promedio igual o superior a 3.5 a nivel anual.
- Reducir la tasa de reseñas negativas en un 10% respecto al año anterior.

## METODOLOGÍA

Se optó por emplear la metodología de tipo Scrum, con reuniones periódicas de 2 o 3 veces por semana y comunicación constante a través de plataformas colaborativas. Las tareas han sido distribuidas entre los miembros del equipo en pos de un desarrollo ágil mientras que se han monitoreado el correcto cumplimiento de estas por parte de los demás miembros del equipo. El proyecto se dividirá en tres Sprints, con entregas de lo elaborado cada 15 días para ser evaluadas para su aprobación o recibirán retroalimentación, según lo determine el Product Owner. Al final de tercer sprint se entregará el producto final al PO.

Los procesos y tiempos del proyecto se podrán ver en el siguiente diagrama de Gantt:




## EDA (Análisis Exploratorio de los Datos)

En este informe se podrán ver el estado de los datasets seleccionados antes de su procesamiento, por lo que podremos ver los datos crudos, la presencia de valores faltantes, ver de antemano la presencia de valores outliers (valores atípicos) y encontrarnos con valores que posteriormente van deberán ser corregidos. Además de todo esto, se han elaborado gráficos que ayudarán al entendimiento y compresión de nuestro reporte EDA. Pueden ver el informe aquí. Una vez terminado el análisis exploratorio de los datos se generará un diccionario de datos. Pueden verlo a través del siguiente link.

Los datasets seleccionados para la realización del proyecto son los siguientes:

| Nombre | Proporcionado por |
|---|---|
| Review - Florida | Google Maps |
| Metadata - Sitios | Google Maps |
| Business | Yelp! |
| Tip | Yelp! |
| Review | Yelp! |

## ETL AUTOMATIZADO

Para llevar a cabo una correcta transformación de los datos de manera automatizada se ha optado por el uso de Google Cloud Platform. Con esta herramienta pudimos llevar a cabo el proceso de ETL (Extract, Transform, Load) a cada uno de los datasets. Para este proceso hemos seleccionado cómo nuestro Data Lake (contenedor en el que van a estar subidos los datos crudos) a Cloud Storage y cómo Data Warehouse (contenedor donde se van a encontrar los datos una vez limpios) a BigQuery.

Estas dos herramientas estarán conectadas por un pipeline. En nuestro caso hemos seleccionado a Cloud Function cómo nuestro pipeline. Éste, cuando detecte movimiento en Cloud Storage se activará y transformará a los datasets según sea requerido para luego cargarlos en BigQuery.


## MODELO DE MACHINE LEARNING - SVM

Utilizamos el algoritmo SVR (Support Vector Machine for Regression) del modelo Support Vector Machine para predecir el incremento de las calificaciones de las sucursales de McDonald's, en caso de que se realicen mejoras en áreas específicas.

En primer lugar, accedimos a los datos limpios almacenados en nuestro Data Warehouse, (BigQuery) desde un Notebook integrado en Vertex AI. Obtuvimos las reseñas y calificaciones de usuarios para cada sucursal, organizados en un archivo CSV único.

Luego, realizamos un procesamiento de texto en las reseñas, convirtiéndolas a minúsculas, eliminando caracteres inválidos, tokenizando el texto y eliminando stopwords. Posteriormente, categorizamos las reseñas en grupos de baja y alta calificación, y creamos matrices TF-IDF para cada grupo. Estas matrices fueron concatenadas en una matriz principal utilizada como referencia para el modelo.

Seguidamente, dividimos la matriz en conjuntos de entrenamiento y prueba. Entrenamos el modelo SVR de regresión lineal utilizando los datos de entrenamiento de una sola sucursal, permitiendo que el modelo aprenda los coeficientes y sesgos necesarios para realizar predicciones precisas.

Realizamos predicciones para una sucursal utilizando los datos de prueba y el modelo entrenado, evaluando el modelo mediante el cálculo del error cuadrático medio (RMSE) entre las predicciones y los valores reales de las calificaciones.

Finalmente, al obtener resultados satisfactorios, entrenamos el modelo con cada sucursal y almacenamos cada archivo con el modelo entrenado en un diccionario, utilizando la dirección de la sucursal como clave. Este archivo con el modelo entrenado con cada una de las sucursales, fue almacenado en Cloud Storage para su posterior uso por la API.

## API Y SUS ENDPOINTS

El modelo de Machine Learning que desarrollamos fue implementado en un endpoint creado con FastAPI, junto con otros dos endpoints adicionales para ofrecer funcionalidades específicas.

El primer endpoint recibe como parámetro la dirección exacta de una sucursal y retorna un diccionario con el total de reseñas recibidas, la cantidad de reseñas positivas, neutras y negativas.

El segundo endpoint recibe como parámetro la dirección exacta de una sucursal y retorna una nube de palabras con las palabras más frecuentes en las reseñas negativas (calificación menor a 3).

Y en el tercer endpoint, recibe como parámetro la dirección de una sucursal y retorna un diccionario con la calificación actual para esa sucursal, el incremento predicho por el modelo, el porcentaje del incremento y la calificación con el incremento predicho sumado.

## DASHBOARD

Para la generación de este, hemos utilizado la herramienta de Microsoft PowerBI, que, con el uso de su poderoso motor Power Query, hemos podido sacar diversas conclusiones que hemos respaldado con los gráficos perteneciente al dashboard interactivo final al que podrán acceder desde Streamlit.


El dashboard se compone de la siguiente información:

- Sedes: Visualiza métricas y análisis por sedes para comparar el rendimiento entre ubicaciones.
- KPI: Destaca indicadores clave de rendimiento (KPI) con métricas anuales asociadas para evaluar el desempeño a lo largo del tiempo.
- Rankings: Muestra varios "Top 3" identificando las mejores y peores sedes en cantidad de reseñas positivas y calificación promedio.

## CONCLUSIONES Y RECOMENDACIONES FINALES

El nivel de satisfacción de los clientes ha tenido una tendencia negativa en los años 2013 y 2021.

Para el indicador clave de desempeño (KPI) de aumento de reseñas positivas, el valor objetivo se logró y superó entre los años 2012 y 2019. Sin embargo, en 2020 y 2021, no solo no se alcanzó el objetivo, sino que el número de reseñas positivas disminuyó sustancialmente, aproximadamente en un tercio cada año.

En cuanto al KPI de mantenimiento de la puntuación promedio, el valor objetivo no se alcanzó en ninguno de los años evaluados, aunque estuvo muy cerca, siendo menos de dos décimas por debajo durante 2012

Para el KPI de disminución de la tasa de reseñas negativas, el valor objetivo se logró e incluso se superó en los años 2016 y 2017. En los años 2018 y 2019, aunque no se alcanzó el valor objetivo, se registró una disminución de la tasa de reseñas negativas, con una diferencia de menos de 3 puntos porcentuales.

Hay varias sedes con puntuaciones promedio muy bajas, lo cual repercute en los KPI, por esto es importante realizar acciones que promuevan la mejora en los aspectos que generan estas calificaciones.

En resumen, se observa que, aunque no se han alcanzado los valores objetivo de los KPI en los últimos años, son alcanzables, ya que se ha estado muy cerca de conseguirlos o se han alcanzado en años anteriores. Por lo tanto, es de suma importancia identificar los aspectos que más influyen en la obtención de reseñas negativas, con el fin de establecer estrategias que conviertan estas áreas en oportunidades de mejora que impacten positivamente en la satisfacción del cliente.